Name: ROMEO SARKAR Adm. No.: 20JE0814

Lab Exercise 2.1 Write a PyCUDA program to demonstrate the followings:

- 1. Allocate host and device memories for three matrices A, B, C
- 2. Transfer data of matrices A, B from host to device
- 3. Performance Matrix and Matrix multiplication

```
import pycuda.driver as cuda
import pycuda.autoinit
from pycuda.compiler import SourceModule
import numpy
TILE WIDTH = 2;
MATRIX LEN = 8
mat a = numpy.random.randn(MATRIX LEN,
MATRIX LEN).astype(numpy.float32)
mat b = numpy.random.randn(MATRIX LEN,
MATRIX LEN).astype(numpy.float32)
mat c = numpy.empty like(mat a)
dev a = cuda.mem alloc(mat a.nbytes)
cuda.memcpy htod(dev a, mat a)
dev b = cuda.mem alloc(mat b.nbytes)
cuda.memcpy htod(dev b, mat b)
dev c = cuda.mem alloc(mat c.nbytes)
cuda.memcpy htod(dev c, mat c)
source module = SourceModule("""
 global void tiledMatrixMulKernel(float *mat1, float
*mat2, float *mat3, int width) {
         shared float mds[2][2];
```

```
shared float nds[2][2];
        int bx = blockIdx.x;
        int by = blockIdx.y;
        int tx = threadIdx.x;
        int ty = threadIdx.y;
        int row = by * 2 + ty;
        int col = bx * 2 + tx;
        float pvalue = 0;
        for(int ph = 0; ph < width / 2; ph++) {
                mds[ty][tx] = mat1[row * width + ph * {0}
+ tx];
                nds[ty][tx] = mat2[(ph * 2 + ty) * width
+ col];
                syncthreads();
                for (int k = 0; k < 2; k++) {
                        pvalue += mds[ty][k] *
nds[k][tx];
                  syncthreads();
        mat3[row * width + col] = pvalue;
111111
tiled matrix multiplication function =
source module.get function("tiledMatrixMulKernel")
tiled matrix multiplication function(dev a, dev b, dev c,
MATRIX LEN, block=(1, 1, 1), grid=(MATRIX LEN,
MATRIX LEN, 1))
```

```
cuda.memcpy_dtoh(mat_c, dev_c)

print("Matrix A:")
print(mat_a)
print("Matrix B:")
print(mat_b)
print("Product Matrix:")
print(mat_c)

docker run --runtime=nvidia -v $HOME:$HOME -ti
bryankp/pycuda:latest bash
```